Induced Start Dynamic Sampling for Wafer Metrology Optimization

Induced Start Dynamic Sampling for Wafer Metrology Optimization

Gian Antonio Susto, Marco Maggipinto, Federico Zocco, Seán McLoone

***Abstract*—Metrology, which plays an important role in ensuring production quality in modern manufacturing industries, incurs substantial costs, both in terms of the infrastructure required, and the time needed to perform measurements. In particular, in the semiconductor manufacturing industry, measuring fundamental quantities on different sites of a wafer surface is associated with increased production time. To increase metrology efficiency, a typical strategy is to limit the number of sites measured and to exploit statistical models (soft sensing) to reconstruct the wafer profile. Moreover, for quality reasons, spatial dynamic sampling strategies may be employed to ensure that all regions of a wafer surface are checked periodically during production. In this work, we propose a new sampling strategy, called Induced Start Dynamic Sampling (ISDS), that adapts greedy feature selection algorithms to the spatial dynamic sampling problem such that the number of measured sites at each process run is minimized while achieving good wafer profile reconstruction accuracy and process visibility. The superiority of the proposed strategy with respect to the state-of-the-art is demonstrated using both simulated data and an industrial chemical vapour deposition case study.**

***Note To Practitioners*–In this work we tackle a practical metrology problem encountered in semiconductor manufacturing, namely, the design of a dynamic wafer measurement plan to monitor the accuracy of a process across the whole wafer surface. The measurement plan is called 'dynamic', since, the measurement locations, which are drawn from a candidate set that provides coverage of the whole wafer, change at each process iteration. Our methodology addresses the challenge of finding an optimized trade-off between the number of measurements performed on each wafer and the reconstruction accuracy that can be achieved for the unmeasured areas on the wafer, while at the same time, for quality assurance purposes, ensuring that all locations on a wafer are visited in a finite number of process runs. The major benefit of the methodology is that it can significantly reduce the number of sites that need to be measured on each wafer enabling greater throughput on metrology tools without sacrificing process monitoring and anomaly detection capability.**

***Index Terms*—Chemical Vapor Deposition, Dynamic Sampling, Feature Selection, Industry 4.0, Semiconductor Manufacturing, Virtual Metrology, Wafer Profile Reconstruction**

I. Introduction and Literature Review

THE Industry 4.0 revolution is based on data availability regarding every aspect of business, industrial and manufacturing processes [1], [2]. In this context, metrology has become more relevant than ever in manufacturing [3]. Unfortunately, metrology is generally expensive due to the need for capital expenditure on measurement, data logging and archiving infrastructure, and the cost of operating and maintaining such infrastructure. It also represents a non-added value cycle-time overhead and often becomes a limiting factor in production line throughput. From this perspective, several machine learning-based approaches that seek to reduce metrology steps/costs without sacrificing quality and process monitoring performance (e.g. Virtual Metrology/Soft Sensing [4]–[6] and Dynamic Sampling [7]), have proliferated in recent years.

In particular, in the semiconductor manufacturing industry, one of the most data-intensive manufacturing industries, metrology efficiency is becoming extremely relevant [8], [9]. In semiconductor fabrication, electronic circuits are created through multi-step sequences of chemical and photolithographic processes applied to *wafers*, disks made of semiconducting material. In some of the aforementioned production steps, such as Chemical Vapor Deposition (CVD), the step quality is assessed by measuring a set of, so-called, critical dimensions at several sites on the wafer surface, with consequent time-consuming procedures needed to allow the metrology tool to visit all the required sites. In fact, in the example case of CVD, a process widely employed for altering the chemical composition of wafers, the quality is assessed by measuring the height of the 'deposited' layer [10]. The challenge for CVD and similar spatial processes is measuring a sufficient number of sites to allow the complete wafer profile to be reconstructed. This is increasingly becoming an issue with the move to larger wafer diameters in modern semiconductor fabrication facilities. As stated above, metrology operations are usually extremely resource intensive both in terms of time and money, hence it is usually not practical to obtain high density measurements; instead, sampling locations/instants are reduced to a meaningful set of points that are chosen by exploiting prior knowledge of the process behavior or by means of automatic space filling procedures [11].

Often, as process knowledge increases over time and previously unseen behaviors are detected, new sampling points are added in order to capture important information that was undetectable with the original set of measurements. During this process, the new measurements points are selected without taking into account the information redundancy that they may provide. In such a scenario, it is extremely valuable to perform a data-driven analysis of the problem in order to optimize the measurement plan, limiting the redundancy in the performed metrology operations. Usually, such optimization requires the identification of a subset of the measured locations that provides minimum reconstruction error at the unmeasured

G.A. Susto (corresponding author) and M. Maggipinto are with the Department of Information Engineering, University of Padova, Italy. E-mail: gianantonio.susto@dei.unipd.it.

F. Zocco and S. McLoone are with Queen's University Belfast, Belfast, Northern Ireland, United Kingdom.

sites. This problem is closely related to the problem of feature selection, that has been extensively studied in the field of machine learning, where the goal is to select the most meaningful variables to employ in a prediction problem. In such problems reducing the number of features has the advantage of preventing overfitting and multicollinearity [12] allowing better prediction capabilities.

A widely employed method for dimensionality reduction is Principal Component Analysis (PCA) [13] but it does not provide information on which features are the most relevant since each principal component is a linear combination of all the available variables. Sparse PCA tries to solve this problem by introducing a sparsity constraint on the projection matrix but it does not provide an appropriate feature selection method. Consequently, other methods, both supervised and unsupervised, have been developed that are able to determine a subset of meaningful features. A common supervised feature selection method is the so-called LASSO [14] regression which introduces an L1-norm penalization term into the regression cost function. This has been shown to have the desirable property of providing a sparse solution for the regression coefficients and hence can be tuned to only select the most relevant features. LASSO have been extensively employed in the machine learning community and also in data-driven semiconductor manufacturing technologies [14]–[17]. Other supervised methods have been employed in the semiconductor industry such as Least Angle Regression [18] and Forward Selection Regression [19] where features are added sequentially until the performance improvement does not justify the complexity increment.

In general, metrology plan optimization is unsupervised in nature, in that the requirement is to select the most representative locations based only on the measurements available, without having a specified prediction task as a final goal. In [20] a two step selection method based on k-Nearest-Neighbours is proposed. In the first step, the k nearest features are computed for each feature. Then the features having the most compact subset (determined by the distance from its farthest neighbour) are selected as the most representative ones. In [21] the authors proposed a method called Orthogonal Principal Feature Selection (OPFS) that exploits the effectiveness of PCA in a greedy fashion, sequentially selecting the feature most correlated with the first principal component of the residual matrix. At each iteration, the residual matrix is updated by projecting it on the space orthogonal to the selected feature thus removing its contribution to the correlation at subsequent steps of the algorithm. In [22] a data-driven Sensor Placement methods was proposed based on Frame Potential (FP), an orthogonality measure of the columns of a matrix. FP has subsequently been employed in [23] as a greedy feature selection algorithm for machine learning problems.

In [24] a novel wafer site selection method has been proposed based on Forward Selection Component Analysis (FSCA) [25], an unsupervised extension of Forward Selection Regression that determines the contribution that individual sites make to the variability observed in a process across a set of candidate wafer sites. In this way, it greedily selects the variables that provide the major contribution to the reconstruction of the unmeasured sites via regression models. A different approach have been proposed in [26] for the purpose of Sensor Placement, where the authors assume the measurements at the sensed locations to be jointly Gaussian. Then, a greedy algorithm is proposed to maximize the mutual information between the selected and unselected locations. Hereafter this approach is referred to as Information Theoretic Feature Selection (ITFS).

The aforementioned approaches enable the implementation of a, so-called, *static* sampling strategy, where a subset of sites (with a given cardinality) is selected, given its optimality in reconstructing the unmeasured sites, and remains unchanged for every process iteration. In wafer metrology, where additional measurement sites are often added to detect specific process peculiarities, a static selection such as the one provided by the aforementioned greedy algorithms is not desirable; there is in fact the risk of missing important process information by having sites that are never measured. For this reason, *dynamic* sampling strategies are preferable in an industrial environment. With such strategies the set of measured sites (with a fixed cardinality) is changed at each process iteration in a way that, over a limited number of process iterations, all the available sites are visited. The challenge in designing dynamic sampling strategies is to achieve full coverage of the sites, while minimising the trade-off in reconstruction capabilities.

A dynamic extension of FSCA was proposed in [27], referred to as Sequential Dynamic Sampling (SDS), where after the static selection, each unselected feature is associated with a cluster induced by the selected ones. Then at each process run, one feature is selected sequentially from each cluster guaranteeing a complete span of the available sites after a limited number of process runs. This approach presents some substantial limitations: on the one hand, at each iteration the group of measured sites is selected based on a similarity measure (e.g. correlation [27]) that generated the clusters, but the resulting reconstruction error is not considered. On the other hand, the number of possible combinations of measured sites is typically very large (proportional to the product of the cluster sizes) and there are no guarantees that the selected combination at each iteration is optimal in terms of reconstruction capabilities. As a result of the aforementioned problems, SDS provided poor reconstruction perform overall when compared to its static counterpart. It should be noted that the static sampling solution represents an upper bound on the performance that can be achieved by dynamic sampling approaches.

In this paper, we propose a novel approach to dynamic sampling, called *Induced Start Dynamic Sampling* (ISDS), that provides a natural extension of greedy selection methods to dynamic sampling, and is able to achievable comparable reconstruction accuracy to the static sampling gold standard. In so doing, it consistently outperform SDS [27], the current state-of-the-art approach for wafer spatial dynamic sampling. To demonstrate the efficacy of the ISDS strategy, it is implemented with the four greedy selection algorithms described above, namely, FSCA, OPFS, ITFS and FP, and the resulting dynamic sampling algorithms benchmarked against other dynamic methods proposed in the literature using both

simulated and real industrial case studies.

The remainder of the paper is organized as follows: Section II is dedicated to introducing the aforementioned greedy selection algorithms. Section III describes the proposed ISDS methodology. Section IV introduces the case study datasets, while section V describes the validation experiments conducted and the results obtained. Finally, conclusions and final remarks are presented in Section VI.

II. Methodologies

Data-driven feature selection methods leverage historical data of the process under study in order to select the most relevant variables involved. Usually, the collected data are organized in a so-called *design-matrix* $\mathbf{X} \in \mathbb{R}^{n \times p}$ where n is the number of observations and p is the number of measured variables (features). In the spatial sampling problem at hand, n represents the number of measured wafers and p the number of sites on the wafer surface where the metrology operations take place. The goal is thus to determine an optimal subset of locations S that guarantees the reconstruction of the unmeasured sites with minimum error, which means selecting the most relevant columns of the matrix $\mathbf{X}$ in the sense of providing the most information on the remaining columns. For a thorough explanation of the feature selection methods we introduce the following notation: we denote with V the set of available locations, S the subset of locations selected by a given algorithm and $U = V \setminus S$ the set of unselected sites. Associated with the location sets, we introduce the index sets I_S, I_V and I_U that contain the indexes of the corresponding columns of $\mathbf{X}$.

In the following, the four different greedy feature selection algorithms considered in our work are outlined. Each algorithm provides a set of k selected features. However, in many applications, k is not predefined, but is instead chosen as a trade-off between solution complexity and reconstruction performance. We remark that the solution complexity is, in this context, the number of measured sites as this dictates the time required to perform the measurement of each wafer. This has to be kept to a minimum level to maximize throughput. In these circumstances, a common practice is to initially set $k = p$ to obtain an ordered list of the the available sites: the reconstruction accuracy is then computed for all values of k from 1 to p with cross-validation procedures employed to identify the optimal value.

A. Forward Selection Component Analysis

FSCA, as presented in [24], is a greedy algorithm that sequentially selects the sites on a wafer that explain the highest amount of variance in the data, guaranteeing a good reconstruction of the sites that remain unmeasured. The pseudocode for the algorithm is given in Algorithm 1.

The algorithm takes as input the design matrix $\mathbf{X}$ and the number of components to be selected k. We can identify two main steps in the algorithm:

(i) *the minimization step (line 5)*: Here the site that minimizes the reconstruction error over the dataset by regression on the site measurements is selected. $\hat{\mathbf{X}}(\tilde{\mathbf{x}}_{i^*})$ is thus the reconstruction of $\tilde{\mathbf{X}}$ induced by the column $\tilde{\mathbf{x}}_{i^*}$ of $\tilde{\mathbf{X}}$. If the reconstruction is performed by the Ordinary Least Squares algorithm, the estimation of the unmeasured site values is obtained as:

$$\hat{\mathbf{X}}(\mathbf{x}_{i^*}) = \frac{\tilde{\mathbf{x}}_{i^*}\tilde{\mathbf{x}}_{i^*}^T}{\tilde{\mathbf{x}}_{i^*}^T\tilde{\mathbf{x}}_{i^*}}\tilde{\mathbf{X}}. \tag{1}$$

The reconstruction error is expressed in terms of the Frobenius norm of the difference between the two matrices. The Frobenius norm for a matrix $\mathbf{A}$ is defined as $||\mathbf{A}||_F^2 = \sum_i \sum_j a_{ij}^2$.

(ii) *the deflation step (line 8)*: This removes the contribution of the selected column $\mathbf{x}_{i^*}$ to the estimation of $\tilde{\mathbf{X}}$. This can be seen as a projection of the data on the subspace orthogonal to the selected feature.

Algorithm 1 FSCA

Input: $\mathbf{X}, k$

1: $I_S = \emptyset$
2: $I_U = I_V$
3: $\tilde{\mathbf{X}} = \mathbf{X}$
4: **for** $j = 1 \dots k$ **do**
5: $\quad i^* = \underset{i \in I_U}{argmin}\, ||\tilde{\mathbf{X}} - \hat{\mathbf{X}}(\tilde{\mathbf{x}}_i)||_F^2$
6: $\quad I_S = I_S \cup \{i^*\}$
7: $\quad I_U = I_V \setminus I_S$
8: $\quad \tilde{\mathbf{X}} = \tilde{\mathbf{X}} - \hat{\mathbf{X}}(\tilde{\mathbf{x}}_{i^*})$
9: **end for**
10: **return** I_S

B. Information Theoretic Feature Selection

The ITFS approach has been presented in [26] to solve sensor placement problems in a data-driven fashion where a collection of historical measurements is available. The assumption upon which ITFS is based is that the measurements are distributed as a multivariate-Gaussian distribution; with this hypothesis a formal mathematical expression can be defined for the problem of maximizing the Mutual Information (MI) between the measured and unmeasured locations. By maximizing MI it is possible to identify a subset of measured locations that provides the maximum amount of information about the unmeasured ones, thus providing a good reconstruction of the sensed field.

We recall that for a multivariate-Gaussian random variable $\mathbf{X}$ with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ the entropy depends exclusively on the determinant of the covariance matrix:

$$H(\mathbf{X}) = \frac{1}{2} ln[(2\pi e)^n det(\boldsymbol{\Sigma})]. \tag{2}$$

Under the Gaussian assumption, we can develop a Gaussian Process Regression (GPR) model [28]. Given a set of locations S with the associated vector of output values $\boldsymbol{s}$, the probability distribution of the measurements at new locations U is normal with mean $\boldsymbol{\mu}^*$ and covariance matrix $\boldsymbol{\Sigma}^*$ that can be computed as follows:

$$\boldsymbol{\mu}^* = \boldsymbol{K}(U,S)(\boldsymbol{K}(S,S) + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{s} \tag{3}$$

$$\boldsymbol{\Sigma}^* = \boldsymbol{K}(U,U) - \boldsymbol{K}(U,S)(\boldsymbol{K}(S,S) + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{K}(S,U) + \sigma^2\boldsymbol{I}, \tag{4}$$

where σ is an hyperparameter and $\boldsymbol{K}(\cdot,\cdot)$ is a correlation function that often depends only on the distance (spatial) between two locations. In our case, since multiple measures at the same location are available, the correlation function can be determined in a data-driven way; given a design matrix $\mathbf{X}$ of size $n \times p$ where the columns are centered around zero, the estimate of the correlation function at the p locations is:

$$\mathbf{\Sigma} = \frac{1}{N-1}\mathbf{X}^T\mathbf{X}. \tag{5}$$

Given a set of k selected locations S and a set of unselected locations U, the MI maximization problem can be expressed as:

$$\begin{aligned} S^* &= \underset{S\subset U:\,|S|=k}{argmax}\ MI(S,U) \\ &= \underset{S\subset U:\,|S|=k}{argmax}\ H(S) - H(S|U). \end{aligned} \tag{6}$$

For the Gaussian assumption the measurements at S and U are distributed as a multivariate-Gaussian with zero-mean and covariance matrix $\mathbf{\Sigma}_S$ and $\mathbf{\Sigma}_U$ that can be estimated as stated in Eq. (5). Since the maximization problem expressed in Eq. (6) is NP-complete [26], a greedy algorithm is proposed whereby sensors are added sequentially such that at a given iteration, the next sensor v^* which provides the maximum increase in MI [26] is chosen:

$$\begin{aligned} \Delta_{MI} &= I(S\cup v;U) - I(S;U\cup v) \\ &= H(v|S) - H(v|U). \end{aligned} \tag{7}$$

For the GPR model, the measurement at v is Gaussian with conditional distribution $v|S \approx \mathcal{N}(\mu_S^*, \Sigma_S^*)$ and $v|U \approx \mathcal{N}(\mu_U^*, \Sigma_U^*)$, where Σ_S^* and Σ_U^* can be computed using Eq. (4). In particular:

$$\Sigma_S^* = \sigma_v - \mathbf{\Sigma}_{Sv}^T\mathbf{\Sigma}_S^{-1}\mathbf{\Sigma}_{Sv}, \tag{8}$$

where $\sigma_v = \frac{1}{N-1}\bar{\mathbf{X}}_{I_v}^T\bar{\mathbf{X}}_{I_v}$, $\mathbf{\Sigma}_{Sv} = \frac{1}{N-1}\bar{\mathbf{X}}_{I_S}^T\bar{\mathbf{X}}_{I_v}$, $\sigma = 0$ and the vector $\bar{\mathbf{X}}_{I_v}$ is equal to the column of $\mathbf{X}$ corresponding to the location v normalized to have zero mean. Σ_U^* is obtained in a similar fashion. It is worth remarking that, in this case, the covariance matrix reduces to a scalar since a single new location is considered. Under the Gaussian assumption the entropy difference expressed in Eq. 7 is a monotonically increasing function of $\frac{\Sigma_S^*}{\Sigma_U^*}$ [26], hence the maximum increase in MI is provided by the location that maximizes $\frac{\Sigma_S^*}{\Sigma_U^*}$. The selection algorithm can be expressed as follows:

Algorithm 2 ITFS

Input: $\mathbf{X}, k$
1: $S = \emptyset$
2: $U = V$
3: **for** $j = 1 \dots k$ **do**
4: $\quad v^* = \underset{v\in U}{argmax} \frac{\sigma_v - \mathbf{\Sigma}_{Sv}^T\mathbf{\Sigma}_S^{-1}\mathbf{\Sigma}_{Sv}}{\sigma_v - \mathbf{\Sigma}_{Uv}^T\mathbf{\Sigma}_U^{-1}\mathbf{\Sigma}_{Uv}}$
5: $\quad S = S \cup \{v^*\}$
6: $\quad U = V \setminus S$
7: **end for**
8: **return** $I_S = indexes(S)$

C. Orthogonal Principal Feature Selection

OPFS [21], as summarized in Algorithm 3, is a simple method of performing greedy feature selection based on exploiting the effectiveness of PCA at finding the directions that explain the highest variance in the dataset. OPFS operates by sequentially selecting the feature that is most correlated with the first Principal Component ($\boldsymbol{p}_1$) of the deflated matrix at each iteration. The algorithm can be divided into three main steps:

1) First PC computation - here the PC associated with the largest eigenvalue of the correlation matrix $\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}$ is computed.
2) Feature selection - the feature most correlated with the first PC is selected.
3) Deflation step - the contribution of the selected feature is removed from the residual matrix, in a similar fashion to FSCA.

Algorithm 3 OPFS

Input: $\mathbf{X}, k$
1: $I_S = \emptyset$
2: $I_U = I_V$
3: $\tilde{\mathbf{X}} = \mathbf{X}$
4: **for** $j = 1 \dots k$ **do**
5: $\quad \boldsymbol{p}_1 =$ First principal component of $\tilde{\mathbf{X}}$
6: $\quad i^* = \underset{i\in I_U}{argmax}\, \rho(\tilde{\mathbf{x}}_i, \boldsymbol{p}_1)$
7: $\quad I_S = I_S \cup \{i^*\}$
8: $\quad I_U = I_V \setminus I_S$
9: $\quad \tilde{\mathbf{X}} = \tilde{\mathbf{X}} - \hat{\mathbf{X}}(\tilde{\mathbf{x}}_{i^*})$
10: **end for**
11: **return** I_S

The correlation at step 2 is expressed in terms of the Pearson correlation coefficient. Given two vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ the Pearson correlation coefficient ρ is defined as

$$\rho(\mathbf{x}_i, \mathbf{x}_j) = \frac{\mathbf{x}_i^T\mathbf{x}_j}{\sqrt{(\mathbf{x}_i^T\mathbf{x}_i)(\mathbf{x}_j^T\mathbf{x}_j)}}, \tag{9}$$

where the vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ are assumed to have zero mean.

D. Frame Potential

In [22] a data-driven sensor placement method was presented that maximizes an objective function defined by the Frame Potential (FP) of a matrix formed by the measures at the selected locations. The Frame Potential of a matrix $\mathbf{X}$ is defined as:

$$FP(\mathbf{X}) = \sum_{i,j\in I_V} |\langle \mathbf{x}_i, \mathbf{x}_j \rangle|^2. \tag{10}$$

The authors proved that thanks to the sub-modularity property of FP, optimization of the FP based cost function guarantees a near-optimal mean squared error (MSE) solution for linear inverse problems. In [23] the same metric has been employed as the basis for a greedy feature selection algorithm for machine learning problems, but it is noted that in this more general setting MSE performance guarantees are not available.

The FP provides a measure of the orthogonality of the columns of matrix $\mathbf{X}$. The goal of FP based feature selection is thus to find a sub-matrix $\mathbf{X}_{I_S}$ of the design-matrix $\mathbf{X}$ with minimum FP or, equivalently, with maximum orthogonality of the columns. The combinatorial nature of the minimization problem necessitates the use of a greedy heuristic approach in order to obtain a solution in polynomial time. Ranieri et. al. [22] also show that normalizing the columns of $\boldsymbol{X}$ so that they have unit norm is a useful pre-processing step. For this reason, in the following we will assume that the columns of $\boldsymbol{X}$ have unit norm. The FP-based procedure is given in Algorithm 4.

Algorithm 4 FP

Input: $\mathbf{X}, k$

$I_S = \emptyset$
$I_U = I_V$
for $j = 1 \dots k$ **do**
  $i^* = \underset{i \in I_U}{argmin} FP(\mathbf{X}_{I_S \cup \{i\}})$
  $I_S = I_S \cup \{i^*\}$
  $I_U = I_V \setminus I_S$
end for
return I_S

III. Dynamic Sampling

The methodologies detailed in Section II represent approaches to tackling the site selection problem that are derived from different, but mathematically rigorous points of view. However, all the methodologies provide a static selection of measured sites which limits their applicability if faced with previously unseen process behavior. A dynamic approach would provide considerable advantages in such a scenario, even at the expense of an increase in the overall reconstruction error. Here, we propose a dynamic framework based on greedy spatial sampling algorithms that guarantees that all sites are visited in a limited amount of time while minimizing the impact on wafer profile reconstruction accuracy.

A. Induced Start Dynamic Sampling (ISDS)

Greedy selection algorithms compute an approximate solution to a combinatorial optimization problem by sequentially selecting the sites that provide the greatest improvement in the cost function. This characteristic is well suited to being exploited in a dynamic framework that, while providing a sub-optimal solution with respect to the original algorithm in terms of reconstruction error, is able to vary the sites selected at each iteration, guaranteeing full coverage of all sites after a finite number of iterations. We remark that this characteristic is particularly appealing for quality reasons in an industrial environment. The rationale is to force the algorithm to start from an initial (not visited at previous iterations) set of q sites indexed by I_{start}, provided as input, and then proceeding with the classical behavior. In the case of FSCA, the pseudo-code for the modified version of the algorithm is reported in Algorithm 5. The extension to the other greedy algorithms is straightforward and based on the same principle.

Algorithm 5 Induced Start FSCA

Input: $\mathbf{X}, k, I_{start}$

$I_S = I_{start}$
$I_U = I_V \setminus I_S$
$\tilde{\mathbf{X}} \leftarrow \mathbf{X} - \hat{\mathbf{X}}(\mathbf{X}_{I_S})$
for $j = 1 \dots k - q$ **do**
  $i^* = \underset{i=1 \dots p}{argmin} ||\tilde{\mathbf{X}} - \hat{\mathbf{X}}(\tilde{\mathbf{x}}_i)||_F^2$
  $I_S = I_S \cup \{i^*\}$
  $I_U = I_V \setminus I_S$
  $\tilde{\mathbf{X}} = \tilde{\mathbf{X}} - \hat{\mathbf{X}}(\tilde{\mathbf{x}}_{i^*})$
end for
return I_S

With the induced start version of the algorithms it is possible to perform ISDS as follows (Algorithm 6):

1) Take as the starting point a set of $q < k$ sites that have not been previously measured and apply the modified version of the chosen greedy algorithm (i.e. Algorithm 5) to select the remaining $k - q$ sites. Save the resulting k sites as the current measurement plan.
2) If all the sites have been measured at least once, then stop, otherwise repeat step one to generate a new measurement plan.

The outcome of this process is a sequence of k-site measurement plans $I_S^{(t)}$, $t = 1 \dots m$, where $m \leq \lceil \frac{p-k}{q} \rceil + 1$. The measurement plan for the i^{th} production wafer is then defined as $I_S^{(i \bmod m)}$. Hence, by design, at least q previously unmeasured locations are guaranteed to be included in each measurement plan. This enables direct control of the trade-off between reconstruction accuracy and the maximum number of production runs m needed to visit all sites. The least impact on reconstruction accuracy occurs when only one new site is added at each iteration (i.e. when $q = 1$), in which case the upper bound on m is $p - k + 1$.

Algorithm 6 Induced Start Dynamic Sampling

Input: $\mathbf{X}, k, q$

$I_S = \emptyset$
$I_U = I_V \setminus I_S$
$t = 1$
while $I_V \setminus I_S ! = \emptyset$ **do**
  $I_{start} =$ select a subset of $min(q, |I_U|)$ sites from I_U
  $I_S^{(t)} = Algorithm\_5(\mathbf{X}, k, I_{start})$
  $I_S = I_S \cup I_S^{(t)}$
  $I_U = I_V \setminus I_S$
  $t = t + 1$
end while
return $I_S^{(1)} \dots I_S^{(m)}$

Rather than cycling through the same measurement plans every m production wafers, it is also possible to generate a new set of measurement plans for each new cycle by re-running Algorithm 6 with a different initial I_{start} set for each cycle. However, this would add substantially to the complexity of ISDS and its implementation, hence cycling through a finite set of plans is preferable in practice. We remark that, given

Algorithm 7 Minimum Correlation Induced Start

Input: $\mathbf{X}, k, q, I_U, I_S$
$I_{start} = \emptyset$
for $k = 1 \cdots q$ **do**
  $i^* =$ index of the site in I_U that is least correlated with the sites in I_S
  $I_{start} = I_{start} \cup i^*$
  $I_S = I_S \cup i^*$
  $I_U = I_U \setminus i^*$
end for
return I_{start}

Algorithm 8 Optimum MSE Induced Start

Input: $\mathbf{X}, k, q, I_U$
if $|I_U| \leq q$ **then**
  return I_U
end if
$minError = inf$
$C = combinations(I_U, q)$
for c ***in*** C **do**
  $I_S = Algorithm\_5(\mathbf{X}, k, c)$
  if $||\tilde{\mathbf{X}} - \hat{\mathbf{X}}(\mathbf{X}_{I_S})||_F^2 < minError$ **then**
    $I_{start} = c$
    $minError = ||\tilde{\mathbf{X}} - \hat{\mathbf{X}}(\mathbf{X}_{I_S})||_F^2$
  end if
end for
return I_{start}

an initial set of sites I_{start} all the greedy algorithms are deterministic, hence, the final set of selected locations does not vary with different executions.

ISDS provides designers with the freedom to decide how to select the set of starting sites I_{start}, and this can be exploited to meet production needs. The simplest approach is to randomly select a set of q unmeasured sites at each iteration. The option also exists of choosing the optimal greedy search selected sites for the first measurement plan ($t = 1$) and then introducing the random selection from $t = 2$ onwards. In this way, the best reconstruction capabilities are retained for the first wafer thanks to the optimal subset of sites being measured, while the following iterations span the entire wafer surface as required by dynamic selection.

A more systematic approach is to employ a temporal or spatial similarity metric to select, as the starting point for the new measurement plan, the site that is most different from the ones selected in previous measurement plans. For example, the least correlated site could be selected or the site that is located furthest from the previously selected sites. The pseudo code for the minimum correlation site selection approach is given in Algorithm 7.

A priori process knowledge or monitoring requirements can also be used to guide the selection of I_{start}. For example, it may be desirable to include a site located at the centre of the wafer and/or at the edge, at every iteration.

Alternatively, for a given q, we can determine the optimum I_{start} with respect to the reconstruction error and/or cycle length m by systematically evaluating all possible I_{start} sets as explained in Algorithm 8. If an exhaustive search is computationally intractable an approximate solution can be obtained using stochastic optimisation techniques.

B. Linear Regression for Profile Reconstruction

Having measured the k selected sites on each wafer it is necessary to estimate the values at the $p - k$ unselected sites in order to reconstruct the wafer profile. This can be achieved by using the measured sites as regressors and estimating prediction models for each of the unmeasured sites using the historical data. While linear and nonlinear regression approaches can be employed, linear models have proven to be sufficient in practice [27], and with much lower complexity are the preferred option. Given a set of input values $\mathbf{v}_i \in \mathbb{R}^{p\times 1}$, and a target output variable y_i, $i = 1 \ldots n$, organized in an $n \times p$ design matrix $\mathbf{X} = [\mathbf{v}_1^T \ldots \mathbf{v}_n^T]^T$ and an $n \times 1$ output vector $\mathbf{y}$, linear regression estimates a model of the form $f(\mathbf{v}) = \boldsymbol{a}^T\mathbf{v} + b$ parametrized by $\boldsymbol{a} \in \mathbb{R}^{p\times 1}$ and $b \in \mathbb{R}$, to predict y such that the prediction error expressed in terms of the Mean Squared Error (MSE) over the dataset is minimised. The optimal model parameters are given by

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{a},b}{argmin} \frac{1}{2} \sum_{i=1}^{n} ||\boldsymbol{a}^T\mathbf{v}_i + b - y_i||_2^2, \tag{11}$$

where $\boldsymbol{\theta} = [\boldsymbol{a}^T \ b]^T$, or equivalently in matrix form as

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{argmin} \frac{1}{2} (\mathbf{X}\boldsymbol{\theta} - \mathbf{y})^T(\mathbf{X}\boldsymbol{\theta} - \mathbf{y}). \tag{12}$$

Here, matrix $\mathbf{X}$ is assumed to be augmented with a unitary column in order to account for the offset term in the vector $\boldsymbol{\theta} \in \mathbb{R}^{p+1}$. The solution to Eq. 12, referred to as the least squares solution, is given by

$$\boldsymbol{\theta}^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \tag{13}$$

Given a linear model with parameters $\boldsymbol{\theta}_j^*$ $j \in I_U$ for each unmeasured site, then, at a new process run, the unmeasured sites v_j $j \in I_U$ can be estimated from the new measures at the selected sites organized in a vector $\mathbf{v}_S \in \mathbb{R}^{k\times 1}$ using the equation

$$\hat{v}_j = \boldsymbol{\theta}_j^{*T}\mathbf{v}_S. \tag{14}$$

We note that a natural extension of linear regression is the use of ridge regression or other regularization methods [12] which help to address overfitting and ill-conditioning problems. However, the greedy site selection algorithms inherently generate well conditioned regressor matrices with regard to wafer profile reconstruction model estimation, hence the use of such extensions is not justified in this instance [29].

IV. Dataset description

In this Section we present the different case study datasets used to test the effectiveness of the proposed ISDS approach. In particular, we have considered a real industrial case study provided by an industrial partner and a simulated case study where the data have been artificially created by means of a mathematical model of the wafer surface.

(i) Industrial case study [Industrial] [27] - This is based on historical metrology data acquired from a CVD process where wafer thickness has been measured at a set of 50 locations (sites) across the wafer surface for a set of $n = 316$ process runs acquired over a time interval of several weeks. Fig. 1 shows example wafer profiles from this dataset. Considerable wafer-to-wafer variability is noticeable. The measurement locations used are reported in Fig. 3(a). For confidentiality reasons, all the reported values have been normalized.

(ii) Simulated case study [RBF] [27] - In this case study the wafer profiles have been artificially generated as a linear combination of Gaussian Radial Basis Functions (RBF) defined on the unit radius disk centered at the origin with additive Gaussian measurement noise. The model for the wafer height at coordinates x, y is given by

$$z(x,y) = \sum_{i=1}^{N_g} h_i e^{\left(\frac{(x-c_{x_i})^2+(y-c_{y_i})^2}{S_f^2}\right)} + \epsilon, \qquad (15)$$

where $h_i \sim \mathcal{N}(0,1)$, $c_{x_i}, c_{y_i} \sim U(-1,1)$ and $\epsilon \sim \mathcal{N}(0, 0.02)$. The number of RBF functions N_g and spreading factor S_f are used to control the smoothness of the resulting wafer profiles and essentially define the spatial variability of the simulated process. Fig. 2 shows some examples of synthetic wafer profiles generated by this model when $S_f = 0.6$ and $N_g = 100$.

The selection of the simulated dataset is motivated by its use in [27] and facilities a fair comparison with the state-of-the art method, SDS. Moreover, in the interest of reproducibility of results, the instance of the dataset used in the paper has been made available online[1]. Fig. 3(b) shows the 50 randomly selected locations where the simulated metrology operations are performed. The locations of the sites on the disk are selected at random while imposing a constraint on the minimum distance between sites on the wafer.

V. Results

In this Section we provide a performance comparison of the proposed ISDS based methods. In order to obtain a statistically robust performance evaluation, a Monte Carlo Cross-Validation procedure has been employed [12] where the datasets are split into training and test sets with a predefined size ratio; then, the model is trained on the training data and its performance evaluated on the previously unseen test data. This process is repeated 1000 times and the results are averaged. Specifically, given the number of available observations N, we create the test set randomly sampling $N_{test} = rN$ observations with $r = 0.35$. In this way, 65% of the data are employed for training and 35% for testing.

The reconstruction performance is reported in terms of the average Mean Squared Error (MSE) over the $p-k$ unmeasured sites. For a single measurement y, the MSE is computed as

$$MSE = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} (\hat{y}_i^{(test)} - y_i^{(test)})^2, \qquad (16)$$

[1]Available at https://gitlab.dei.unipd.it/dl_dei/ISDS-Data

where $\hat{y}$ is the target value predicted by the model and y is the real value. As the purpose of ISDS is to guarantee cyclic measurement of all sites on a wafer surface, it is also necessary to compare the *cycle time* achieved with each method, i.e. the number of wafers processed before all sites are visited. In order to provide a comprehensive performance comparison the following methods, as proposed in [27], will be included in the comparison:

- Random Static - a single set of k sites is randomly selected and employed for all wafers.
- Random Dynamic - sites are randomly ordered and then visited sequentially, k sites at a time.
- Conservative Dynamic Sampling (CDS) - $k-1$ sites are selected according to FSCA with the remaining site selected randomly from the unselected sites.
- Sequential Dynamic Sampling (SDS) - a cluster based method, where a set of initial sites is selected according to FSCA and then the remaining sites are assigned to a cluster generated by each one of the selected sites according to a correlation based similarity metric. For a more detailed description of SDS we invite the reader to refer to [27].

As a baseline for method evaluation, we adopt the simplest implementation of ISDS, with $q = 1$ and I_{start} selected randomly. For a given algorithm the ISDS implementation will be denoted by '<algorithm name>-IS'. For comparision purposes we also include the results for static sampling with each of the greedy methods. While these have a cycle time of infinity they provide a useful reference as the lower bound for the achievable MSE with dynamic sampling.

A. Reconstruction performance

The reconstruction error achieved with the linear regression models is a direct measure of the quality of the sites selected by the algorithms. Figure 4 shows the reconstruction error as a function of the number of selected sites k for the various static and dynamic sampling methods considered in the paper when applied to both the industrial and simulated case studies. The MSE for selected values of k is also reported in Table I for the industrial dataset. It is immediately apparent that FSCA provides the best performance among the static methods. Similarly, FSCA-IS provides the best performance among the dynamic methods. OPFS-IS provides similar performance to FSCA-IS with both methods consistently outperforming SDS. FP based methods tend to be significantly outperformed by the other methods, suggesting that the FP metric does not identify appropriate sites in terms of reconstruction performance [23]. ITFS-IS achieves similar performance to FSCA-IS for the industrial case study but is considerably poorer on the simulated dataset, showing limited general applicability of the method. The performance of CDS falls between FSCA-IS and SDS.

Figure 4 (e) and (f) provide a comparison between the best performing static method, FSCA, the state-of-the-art dynamic sampling method, SDS, and the best performing ISDS method, FSCA-IS. As expected, when k is small FSCA performs better than FSCA-IS, but as k increases, the difference in reconstruction accuracy quickly becomes negligible. A similar pattern

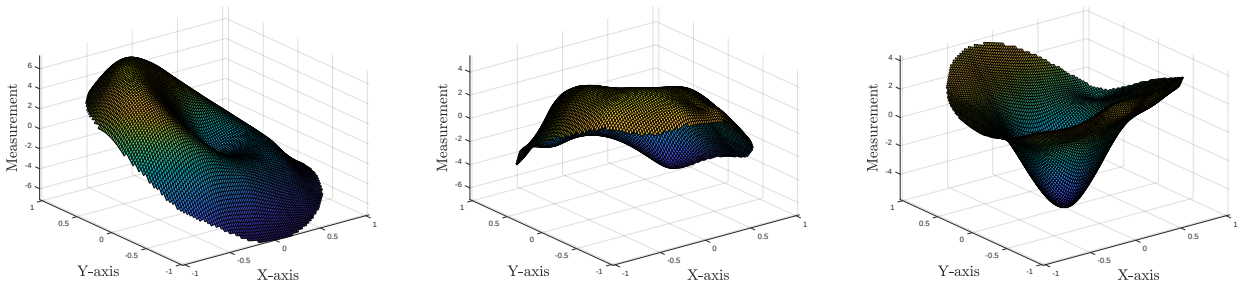


Figure 1. [Industrial] Wafer profiles from an industrial CVD process

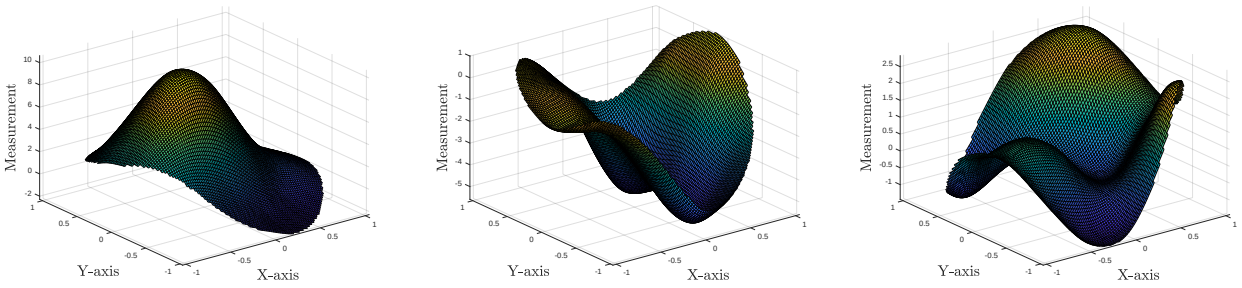
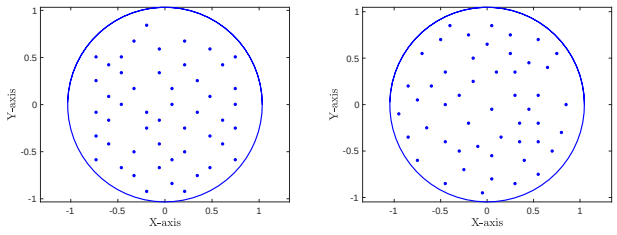


Figure 2. [RBF] Wafer profiles from the simulated dataset for $S_f = 0.6$ $N_g = 100$



(a) Industrial dataset (b) Simulated dataset

Figure 3. Measurement sites locations on the wafer surface for both the industrial and simulated case studies

occurs for SDS, but with much larger MSEs, and a much larger k before performance differences become negligible.

Figure 5 shows the distribution of the MSE over 1000 MCCV cycles for each sampling method when applied on the industrial case study with $k = 7$ sites selected. Here it is evident how the random methods and FP are outperformed by the others approaches. Among the dynamic methods, FSCA-IS and OPFS-IS present similar performance as previously noted.

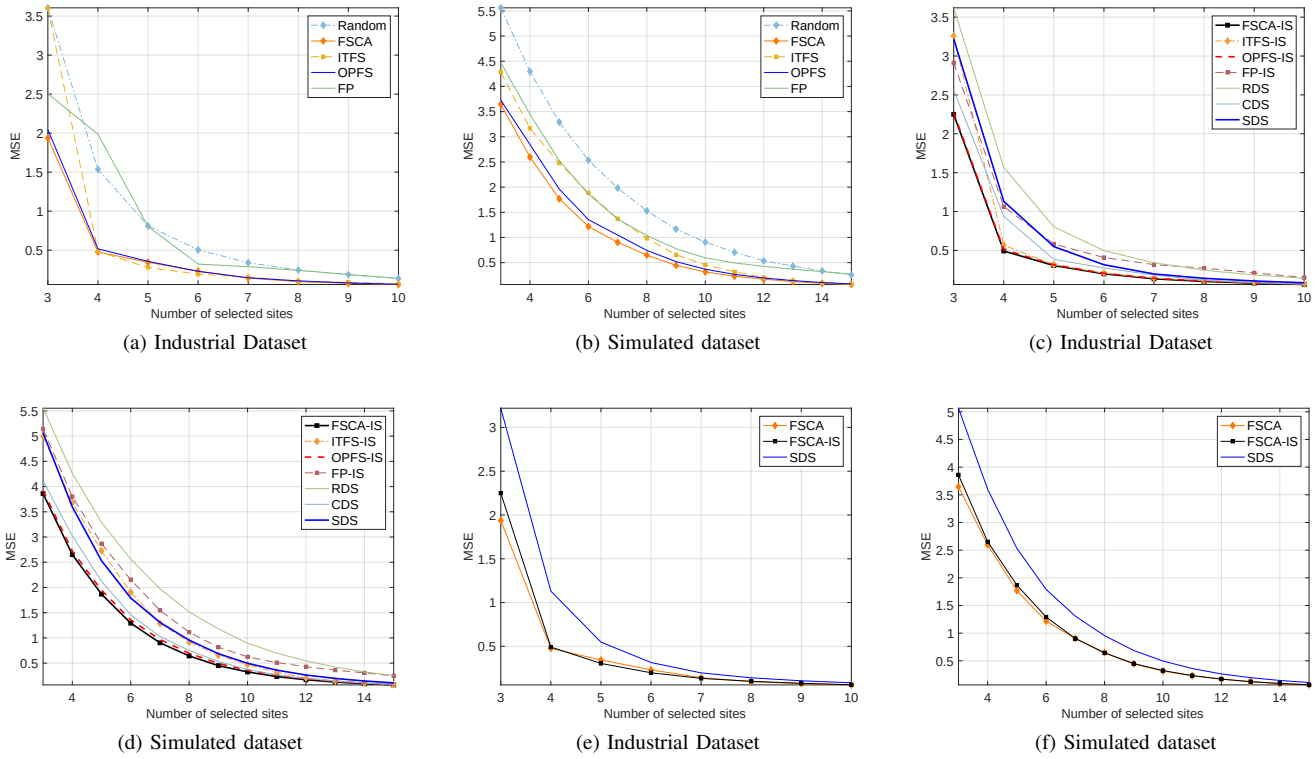
B. Cycle time performance

By employing dynamic methods it is possible to cover sequentially the entire set of available sites by including, at each process run, previously unvisited measurement locations (Figure 6). This of course causes a decrease in wafer profile reconstruction performance due to the sub-optimal measurement set used at each iteration. With the objective of entire wafer coverage, an important factor that determines the quality of the dynamic sampling methods is the cycle length, i.e. the number of iterations that it takes to visit the complete set of available locations. Figure 7 is a plot of the median cycle duration over 1000 MCCV iterations as a function of the number of selected sites k for each of the dynamic sampling methods considered.

Among dynamic sampling methods, CDS represents the worst case scenario in terms of cycle time with only a single new site visited at each iteration, while RDS represents the best case scenario with k new sites visited. Hence, these methods provide upper and lower bounds on achievable cycle time with the other dynamic methods.

FP-IS has the longest cycle duration among the ISDS methods. Observing the number of times that a particular location is measured during a single production cycle of m wafers (Figure 8), it is clear that this arises because FP-IS tends to visit a small subset of sites frequently, and consequently, it takes longer for the unselected locations to be included in the metrology process. The remaining ISDS methods have similar behaviour, providing a good compromise between performance and cycle duration. In particular, FSCA-IS provides the best prediction capabilities combined with a cycle duration that is similar to the other induced methods.

Overall, the cluster based SDS method yields the best cycle duration, but, as previously noted it has much poorer MSE performance than FSCA-IS and OPFS-IS. Significantly, unlike SDS, the ISDS methods provide the freedom to choose the number of initial sites q at each iteration. This parameter can be used to trade-off MSE performance for reduced cycle duration making ISDS a versatile choice for the problem at hand, as will be illustrated in the next section.



(a) Industrial Dataset (b) Simulated dataset (c) Industrial Dataset

(d) Simulated dataset (e) Industrial Dataset (f) Simulated dataset

Figure 4. MSE as a function of the number of measured sites for (a)(b) static methods, (c)(d) dynamic methods and (e)(f) the best static method compared with best dynamic method.

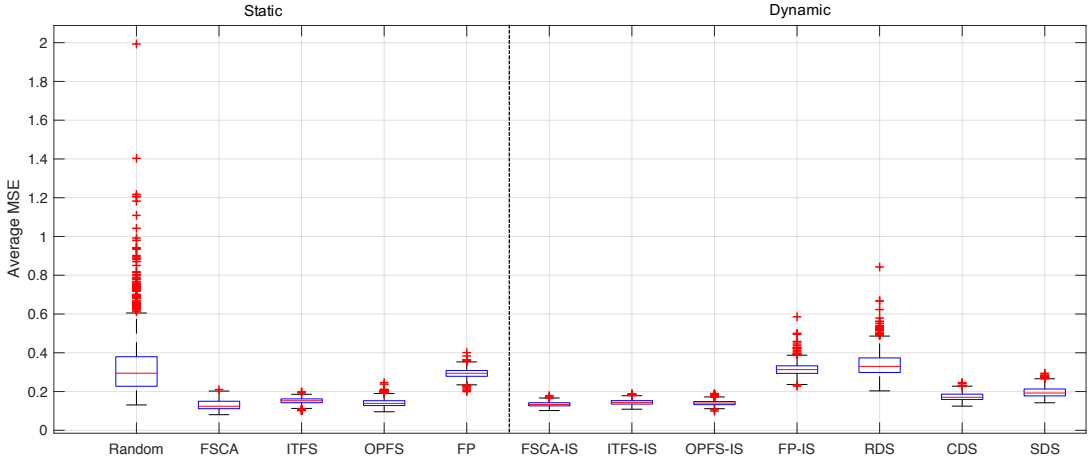


Figure 5. Boxplot of the MSE distribution for wafer sampling method over 1000 MCCV cycles for $k = 7$.

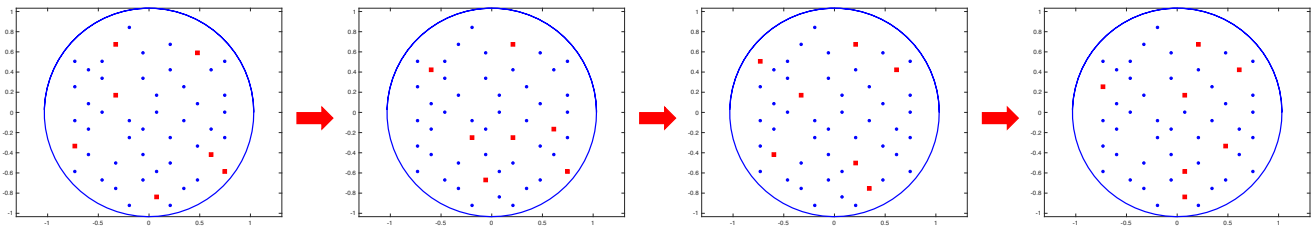
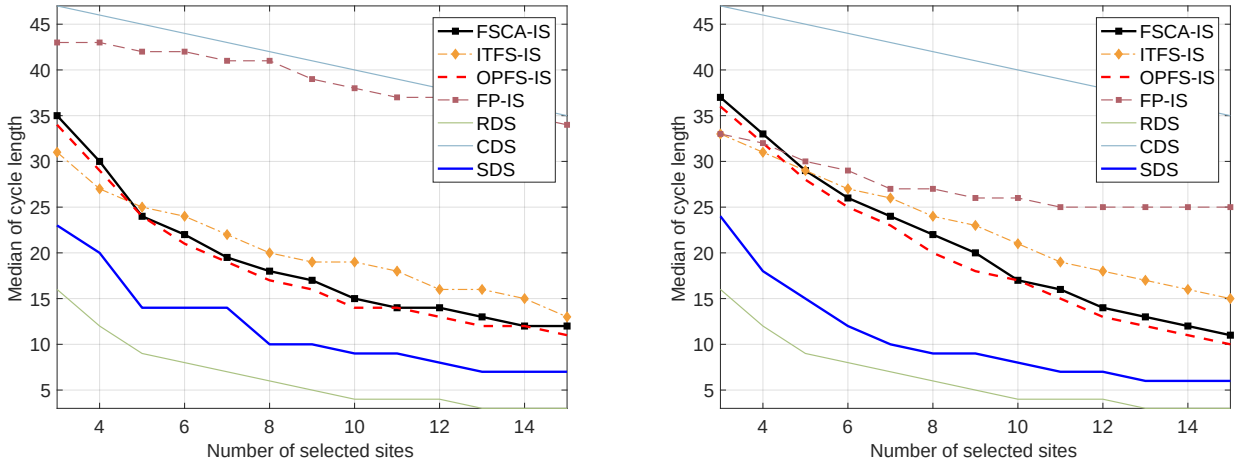


Figure 6. Sample evolution of the measured locations (red) for the FSCA-IS method with $k = 7$.

Table I
MSE AS A FUNCTION OF THE NUMBER OF SELECTED SITES

		MSE			
		$k = 3$	$k = 5$	$k = 7$	$k = 9$
Static	Random	3.60 ± 1.04	0.81 ± 0.54	0.34 ± 0.18	0.19 ± 0.09
	FSCA	**2.03 ± 0.20**	0.34 ± 0.03	**0.13 ± 0.03**	**0.07 ± 0.01**
	ITFS	3.52 ± 0.73	**0.28 ± 0.03**	0.15 ± 0.02	0.08 ± 0.01
	OPFS	2.14 ± 0.30	0.36 ± 0.03	0.14 ± 0.02	0.08 ± 0.00
	FP	2.51 ± 0.27	0.78 ± 0.11	0.29 ± 0.02	0.20 ± 0.04
Dynamic	FSCA-IS	**2.25 ± 0.20**	**0.31 ± 0.02**	**0.13 ± 0.01**	**0.08 ± 0.01**
	ITFS-IS	3.26 ± 0.33	0.32 ± 0.03	0.14 ± 0.01	0.09 ± 0.01
	OPFS-IS	2.27 ± 0.21	0.32 ± 0.03	0.14 ± 0.01	0.08 ± 0.01
	FP-IS	2.91 ± 0.30	0.58 ± 0.07	0.32 ± 0.03	0.21 ± 0.02
	RDS	3.62 ± 0.37	0.80 ± 0.18	0.34 ± 0.06	0.18 ± 0.04
	CDS	2.65 ± 0.26	0.40 ± 0.03	0.17 ± 0.02	0.08 ± 0.01
	SDS	3.22 ± 0.33	0.55 ± 0.08	0.20 ± 0.03	0.11 ± 0.01



(a) Industrial dataset (b) Simulated dataset

Figure 7. Median over 1000 MCCV of the cycle length as a function of the number of measures sites k.

C. Impact of the initial site selection I_{start}

We investigate here the effect that the choice of the starting site, I_{start} $(q = 1)$, has on the performance of ISDS by comparing three possible approaches discussed in section III-A, that is:

(i) **Random** - Choosing the initial site at random.

(ii) **OptMSE** - Choosing the initial site that leads to the best set of measurements in terms of minimizing the wafer profile reconstruction error.

(iii) **Correlation** - Choosing the initial site as the one that is least correlated with the sites that have already been selected in previous iterations.

Fig 9 shows the distribution of reconstruction errors obtained over 1000 MCCV simulations for the two case studies for each of the three initialisation methods. While, 'OptMSE' yields marginally superior results to the other two methods, the differences between the methods are not statistically significant; hence 'Random' selection is a good choice as it has the lowest computation requirements. From a cycle length point of view, Fig. 10 reveals that for this metric "OptMSE" is the worst performing approach while 'Correlation' is the best, and this time the differences are statistically significant. Overall, the minimum correlation based method presents the best trade-off between reconstruction accuracy and cycle time at the expense of a small increase in computational complexity compared to the "Random" method.

D. Dealing with new process behaviour

The introduction of a dynamic sampling approach allows the measuring process to spot previously unseen process behaviour that would go undetected with a static approach. In this section we provide a comparison between FSCA, FSCA-IS and SDS on the industrial dataset where we manually add a new localised process behaviour at a randomly selected location. The perturbation is in the form of a randomly generated RBF surface deviation centered at the selected location. In order to provide a thorough statistical description the comparison is based on a 1000 run Monte Carlo cross-validation study. For each run, a training and a test dataset are created by randomly splitting the dataset into a 222 wafer training set and a 94 wafers test set. The training set is used to train the predictive models, and the test set, modified to include the new process behaviour, is used to evaluate their performance. In order to assess the capacity of each dynamic sampling algorithm to detect the new behaviour, we employ two different approaches based on control limits defined on the training dataset, which

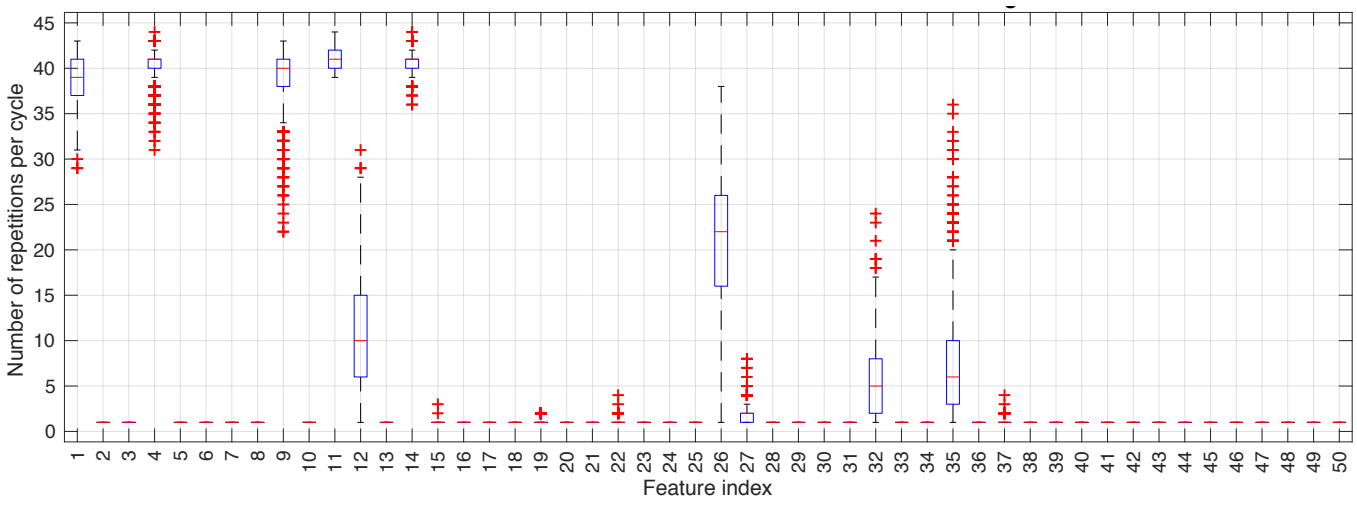
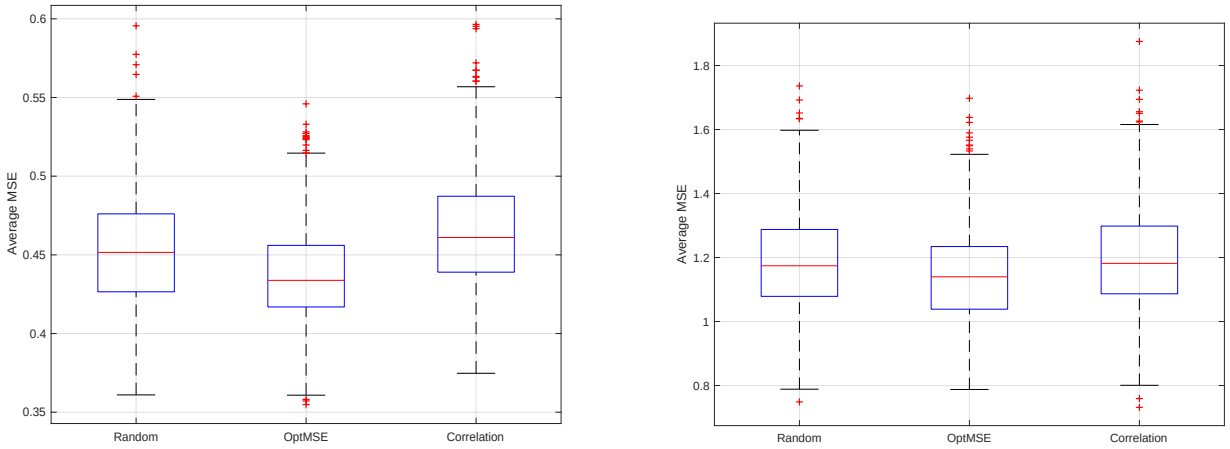
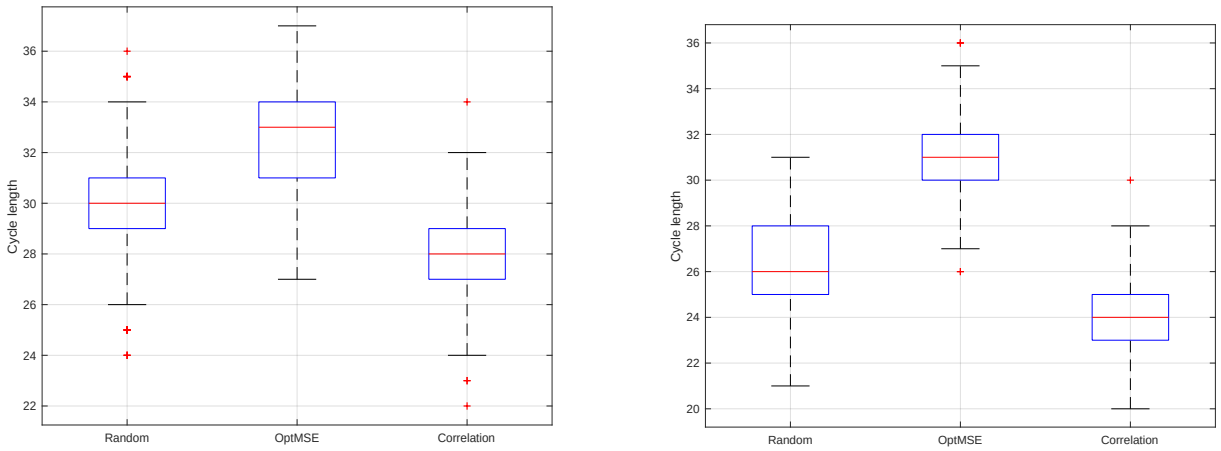


Figure 8. Distribution over 1000 MCCV of the number of times a location is measured during the entire cycle for FP-IS.



(a) Industrial dataset $k = 4$ (b) Simulated dataset $k = 6$

Figure 9. Effect of different choices of the ISDS starting site on the reconstruction error.



(a) Industrial dataset $k = 4$ (b) Simulated dataset $k = 6$

Figure 10. Effect of different choices of the ISDS starting site on the cycle length.

will be referred to as *fault detection* and *anomaly detection* [30]:

- *Fault detection*: a new process behaviour is detected if the measurement at any site (actual or predicted) exceeds the maximum value observed over all sites in the training dataset.
- *Anomaly detection*: a new process behaviour is detected if the measurement at any site (actual or predicted) falls outside the normal range for that site, defined as the mean value +/- three standard deviations of the site measurements in the training dataset.

The difference between the two approaches is that the first employs global control limits while the second employs local control limits, with each site having its own range of normal values. Table II and III present the results obtained with the two different approaches when varying the number of selected sites k and the number of induced start sites q. Two versions of the induced start are considered, namely, the default random site selection approach, FSCA-IS, and the minimum correlation site selection approach, FSCA-IS$^{\text{c}}$. The reported metrics are the detection rate over 1000 MCCV cycles (r_D), the mean number of wafers processed before the anomaly is detected (μ_D), the median of the same quantity (m_D), the reconstruction error without the anomaly included (MSE-) and the reconstruction error with the anomaly included (MSE+). Standard deviation information is also included for μ_D, MSE- and MSE+.

As expected FSCA, the static sampling method, has a detection rate that is much lower than the dynamic methods since it is only able to detect the anomaly when it is occurs at a location that is selected in the measurement plan. The theoretical detection rate for FSCA under these conditions is given by

$$E[r_D] = 100\frac{k}{p}\% \tag{17}$$

Here, $p = 50$, hence $E[r_D] = 2k\%$. The detection rates of each dynamic sampling method approaches 100 % as the number of measured sites per wafer increases, and hence the number of times measurements are taken at the anomaly location over the 94 test wafer validation cycle increases. The theoretical detection rate for the dynamical sampling methods is 100%, but falls short of this here because of the limitations of the experimental setup. Specifically, the stochastic nature of the anomaly generation mechanism employed, combined with the normal wafer variation at each site, means that the magnitude of the resulting anomaly on a given wafer does not always exceed the global or local thresholds limits. This also explains the marginal differences in performance between the global and local detection methods as the local method has inherently greater sensitivity to abnormal variation, and hence is able to detect an anomaly that is greater than the local threshold, but lower than the global threshold.

Taking these factors into account it is clear that the results are consistent for both detection methods. FSCA-IS ($q = 1$) has a smaller reconstruction error compared to SDS but with a longer detection time. However, FSCA-IS provides the freedom to choose the number of starting sites q at each iteration and hence allows a trade-off between reconstruction error and detection time, as demonstrated by the results for FSCA-IS with $q = 2$ and $q = 3$. The superiority of ISDS compared to SDS is evident from the fact that it is possible to reduce the detection time to values that are comparable to SDS while maintaining superior MSE performance. Moreover, SDS achieves the lowest cycle time (and consequently detection time) when the features are evenly distributed among the k clusters. Under these conditions the number of possible combinations of measured sites is at its maximum, with a value greater than or equal to $\lfloor \frac{p}{k} \rfloor^k$, making SDS challenging to implement. This problem does not occur in FSCA-IS where the number of unique site combinations is equal to the cycle length m, which has an upper bound of $\lceil \frac{p-k}{q} \rceil + 1$.

Overall, there is little to choose between the two versions of ISDS considered (FSCA-IS and FSCA-IS$^{\text{c}}$), although, particularly with $q = 1, 2$, FSCA-IS$^{\text{c}}$ does achieve slightly better detection times, which is consistent with the results reported in Fig. 10. FSCA-IS$^{\text{c}}$ has significantly higher computational complexity than FSCA-IS due to the calculation of the correlation matrix (only needs to be done once at the beginning of the algorithm) and the max/min operations that need to be performed at every step. However, determining the dynamic sampling measurement plan is an off-line process, hence the increase in complexity does not have an impact on the on-line performance of the method.

VI. Conclusion

This paper has proposed a novel spatial dynamic sampling methodology for metrology optimization in manufacturing, a problem that is of particular interest where measurement operations are extremely expensive (in terms of time and/or money) and have a significant impact on production costs and yield. The proposed methodology, ISDS, combined with greedy data-driven feature selection techniques, provides an effective framework for minimizing the number of measurements per wafer while enabling reconstruction accuracy to be traded-off against cycle time. The efficacy of the approach has been demonstrated with the aid of both simulated and industrial datasets. A comparison of four greedy-selection algorithms (FSCA, ITFS, OPFS and FP) highlights that ISDS implemented using FSCA (FSCA-IS) yields the best performance, followed closely by the implementation using OPFS (OPFS-IS). Benchmarking of FSCA-IS against other dynamic sampling algorithms, including the state-of-the-art FSCA cluster based SDS algorithm, shows that it yields the best overall performance among the dynamic sampling methods considered in terms of balancing MSE reconstruction performance with cycle time, and hence anomaly detection time. Remarkably, the ISDS approach is optimal in the sense that it achieves similar reconstruction accuracy to the 'static' approaches, which define the lower bound on achievable MSE performance. In addition, the ability to choose the set of starting sites I_{start} at each iteration in ISDS is attractive, as it provides process engineers with the flexibility to incorporate a priori process knowledge or monitoring requirements. In the absence of such requirements selection of I_{start} at random or

Table II
CAPABILITIES OF THE METHODS AT DETECTING A NEW PROCESS BEHAVIOUR USING THE FAULT DETECTION APPROACH (GLOBAL LIMITS)

Number of measured sites	Method name	r_D	μ_D	m_D	MSE-	MSE+
$k = 3$	FSCA	7.30%	12.40 ± 24.32	1	1.91 ± 0.19	12.03 ± 22.73
	FSCA-IS (q=1)	94.30%	22.97 ± 20.01	18	2.13 ± 0.21	12.17 ± 6.06
	FSCA-IS (q=2)	**99.30%**	**15.60 ± 14.62**	**12**	**2.53 ± 0.26**	**15.08 ± 11.44**
	FSCA-IS^c (q=1)	95.30%	20.85 ± 19.45	15	2.17 ± 0.21	12.38 ± 6.49
	FSCA-IS^c (q=2)	99.50%	14.69 ± 13.17	12	2.58 ± 0.27	18.61 ± 16.92
	SDS	**99.80%**	**14.70 ± 13.66**	**11**	**3.04 ± 0.31**	**16.65 ± 6.92**
$k = 5$	FSCA	11.90%	4.04 ± 12.37	1	0.31 ± 0.03	12.15 ± 22.13
	FSCA-IS (q=1)	98.60%	14.41 ± 16.23	10	0.28 ± 0.023	12.22 ± 7.20
	FSCA-IS (q=2)	99.70%	11.35 ± 16.23	8	0.29 ± 0.03	13.55 ± 9.85
	FSCA-IS (q=3)	**99.90%**	**9.49 ± 16.23**	**7**	**0.31 ± 0.03**	**15.43 ± 16.19**
	FSCA-IS^c (q=1)	97.80%	14.28 ± 16.13	9	0.28 ± 0.02	12.21 ± 7.05
	FSCA-IS^c (q=2)	99.70%	10.88 ± 11.27	8	0.29 ± 0.03	13.73 ± 10.21
	FSCA-IS^c (q=3)	99.80%	9.59 ± 9.60	7	0.30 ± 0.03	13.60 ± 8.95
	SDS	**100.00%**	**8.35 ± 7.97**	**7**	**0.50 ± 0.08**	**17.95 ± 16.82**
$k = 7$	FSCA	13.70%	3.82 ± 11.42	1	0.11 ± 0.02	9.19 ± 13.96
	FSCA-IS (q=1)	99.00%	11.46 ± 13.20	7	0.12 ± 0.01	11.31 ± 5.97
	FSCA-IS (q=2)	99.80%	9.48 ± 11.01	7	0.12 ± 0.01	11.63 ± 6.07
	FSCA-IS (q=3)	**99.80%**	**8.24 ± 9.75**	**6**	**0.12 ± 0.01**	**12.70 ± 8.65**
	FSCA-IS^c (q=1)	99.20%	10.92 ± 12.80	7	0.12 ± 0.01	11.06 ± 5.48
	FSCA-IS^c (q=2)	99.80%	9.48 ± 10.92	6	0.12 ± 0.01	11.67 ± 6.04
	FSCA-IS^c (q=3)	99.90%	9.04 ± 10.16	6	0.12 ± 0.01	11.77 ± 6.91
	SDS	**99.80%**	**7.86 ± 8.45**	**6**	**0.17 ± 0.02**	**13.79 ± 7.71**
$k = 9$	FSCA	19.40%	6.57 ± 18.45	1	0.06 ± 0.01	10.25 ± 15.03
	FSCA-IS (q=1)	99.80%	9.02 ± 11.10	6	0.06 ± 0.01	11.19 ± 5.70
	FSCA-IS (q=2)	99.90%	7.89 ± 9.80	5	0.06 ± 0.01	11.90 ± 7.66
	FSCA-IS (q=3)	**99.90%**	**7.79 ± 9.43**	**5**	**0.06 ± 0.01**	**11.69 ± 7.41**
	FSCA-IS^c (q=1)	99.60%	9.03 ± 10.89	6	0.06 ± 0.01	11.25 ± 6.39
	FSCA-IS^c (q=2)	99.80%	7.68 ± 8.70	5	0.06 ± 0.01	11.69 ± 7.15
	FSCA-IS^c (q=3)	100.00%	7.04 ± 7.74	5	0.06 ± 0.01	11.69 ± 7.72
	SDS	**100.00%**	**5.90 ± 6.27**	**4**	**0.09 ± 0.01**	**14.11 ± 7.82**

based on minimum correlation with the already selected sites is recommended.

As future work, we plan to investigate the applicability of ISDS to the optimal design of experiments, a topic that is closely connected to the measurement plan optimization problem described here, and which is relevant in many domains, for example, in medical testing and in chemical manufacturing [31]–[34]. In particular, we note potential parallels that may exist between ISDS and optimal design of experiments with regard to incorporating prior knowledge via initial designs and employing a non-sequential exchange algorithm for targeted design point selection [35]. Finally, we remark that in the dynamic sampling scenario, there is an unavoidable trade-off between reconstruction accuracy and cycle length. In this work we provided a dynamic sampling strategy that achieves MSE performance comparable to static approaches for only a small trade-off in cycle length, and a means to control that trade-off. As future work, we will explore dynamic strategies that can optimize the trade-off while directly accommodating user requirements on maximum allowable cycle length.

REFERENCES

[1] J. Moyne, J. Samantaray, and M. Armacost, "Big data capabilities applied to semiconductor manufacturing advanced process control," *IEEE Transactions on Semiconductor Manufacturing*, vol. 29, no. 4, pp. 283–291, 2016.

[2] M. S. Sivri and B. Oztaysi, "Data analytics in manufacturing," in *Industry 4.0: Managing The Digital Transformation*. Springer, 2018, pp. 155–172.

[3] G. A. Susto, A. Schirru, S. Pampuri, A. Beghi, and G. De Nicolao, "A hidden-gamma model-based filtering and prediction approach for monotonic health factors in manufacturing," *Control Engineering Practice*, vol. 74, pp. 84–94, 2018.

[4] F.-T. Cheng, H.-C. Huang, and C.-A. Kao, "Developing an automatic virtual metrology system," *IEEE Transactions on Automation Science and Engineering*, vol. 9, no. 1, pp. 181–188, 2012.

[5] M. Maggipinto, M. Terzi, C. Masiero, A. Beghi, and G. A. Susto, "A computer vision-inspired deep learning architecture for virtual metrology modeling with 2-dimensional data," *IEEE Transactions on Semiconductor Manufacturing*, vol. 31, no. 3, pp. 376–384, 2018.

[6] P. Zhou, S.-W. Lu, and T. Chai, "Data-driven soft-sensor modeling for product quality estimation using case-based reasoning and fuzzy-similarity rough sets," *IEEE Transactions on Automation Science and Engineering*, vol. 11, no. 4, pp. 992–1003, 2014.

[7] S. Kang and P. Kang, "An intelligent virtual metrology system with adaptive update for semiconductor manufacturing," *Journal of Process Control*, vol. 52, pp. 66–74, 2017.

[8] D. Kurz, C. De Luca, and J. Pilz, "A sampling decision system for virtual metrology in semiconductor manufacturing," *IEEE Transactions on Automation Science and Engineering*, vol. 12, no. 1, pp. 75–83, 2015.

[9] G. A. Susto, "A dynamic sampling strategy based on confidence level of virtual metrology predictions," in *28th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*, 2017, pp. 78–83.

[10] H. Ishii, M. Nagase, N. Ikeda, Y. Shiba, Y. Shirai, R. Kuroda, and S. Sugawa, "A high sensitivity and compact real time gas concentration sensor for semiconductor and electronic device manufacturing process," *ECS Transactions*, vol. 85, no. 13, pp. 1399–1405, 2018.

[11] G. Damblin, M. Couplet, and B. Iooss, "Numerical studies of space-filling designs: optimization of latin hypercube samples and subprojection properties," *Journal of Simulation*, vol. 7, no. 4, pp. 276–289, 2013.

[12] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer series in statistics Springer, Berlin, 2001, vol. 1.

[13] T. J. Rato, J. Blue, J. Pinaton, and M. S. Reis, "Translation-invariant multiscale energy-based pca for monitoring batch processes in semi-

Table III
CAPABILITIES OF THE METHODS AT DETECTING A NEW PROCESS BEHAVIOUR USING THE ANOMALY DETECTION APPROACH (LOCAL LIMITS)

Number of measured sites	Method name	r_D	μ_D	m_D	MSE-	MSE+
k=3	FSCA	5.60%	1.27 ± 0.77	1	1.91 ± 0.19	12.03 ± 22.73
	FSCA-IS (q=1)	98.20%	19.78 ± 16.51	17	2.13 ± 0.21	12.17 ± 6.06
	FSCA-IS (q=2)	**99.80%**	**12.89 ± 10.84**	**11**	**2.53 ± 0.25**	**15.08 ± 11.44**
	FSCA-IS (q=3)	100.00%	10.84 ± 9.17	9	3.37 ± 0.38	24.75 ± 39.13
	FSCA-IS^c (q=1)	97.90%	18.07 ± 16.36	13	2.17 ± 0.21	12.38 ± 6.49
	FSCA-IS^c (q=2)	99.80%	12.43 ± 10.30	11	2.58 ± 0.27	18.61 ± 16.92
	SDS	**99.80%**	**12.62 ± 11.16**	**10**	**3.04 ± 0.31**	**16.65 ± 6.92**
k=5	FSCA	11.20%	1.14 ± 0.48	1	0.31 ± 0.03	12.15 ± 22.13
	FSCA-IS (q=1)	99.60%	11.64 ± 12.04	8	0.28 ± 0.02	12.22 ± 7.20
	FSCA-IS (q=2)	99.80%	9.15 ± 8.99	7	0.29 ± 0.02	13.55 ± 9.85
	FSCA-IS (q=3)	**100.00%**	**7.79 ± 6.99**	**6**	**0.31 ± 0.03**	**15.43 ± 16.19**
	FSCA-IS^c (q=1)	99.50%	11.57 ± 12.45	8	0.28 ± 0.02	12.21 ± 7.05
	FSCA-IS^c (q=2)	99.80%	8.58 ± 7.32	7	0.29 ± 0.03	13.73 ± 10.21
	FSCA-IS^c (q=3)	100.00%	7.95 ± 7.10	7	0.30 ± 0.03	13.60 ± 8.95
	SDS	**100.00%**	**6.89 ± 5.51**	**6**	**0.50 ± 0.08**	**17.95 ± 16.82**
k=7	FSCA	13.00%	1.45 ± 1.12	1	0.11 ± 0.02	9.19 ± 13.96
	FSCA-IS (q=1)	99.90%	8.83 ± 9.15	6	0.12 ± 0.01	11.31 ± 5.97
	FSCA-IS (q=2)	99.90%	7.08 ± 7.64	6	0.12 ± 0.01	11.63 ± 6.07
	FSCA-IS (q=3)	**100.00%**	**6.50 ± 6.56**	**5**	**0.12 ± 0.01**	**12.70 ± 8.65**
	FSCA-IS^c (q=1)	99.80%	8.51 ± 6.56	6	0.12 ± 0.01	11.06 ± 5.48
	FSCA-IS^c (q=2)	100.00%	7.23 ± 7.26	6	0.12 ± 0.01	11.67 ± 6.04
	FSCA-IS^c (q=3)	100.00%	6.84 ± 6.24	6	0.12 ± 0.01	11.77 ± 6.91
	SDS	**100.00%**	**6.35 ± 5.48**	**5**	**0.17 ± 0.02**	**13.79 ± 7.71**
k=9	FSCA	17.90%	2.15 ± 7.78	1	0.06 ± 0.01	10.25 ± 15.03
	FSCA-IS (q=1)	100.00%	7.00 ± 7.46	5	0.06 ± 0.01	11.19 ± 5.70
	FSCA-IS (q=2)	100.00%	5.96 ± 7.46	5	0.06 ± 0.01	11.90 ± 7.66
	FSCA-IS (q=3)	**100.00%**	**5.65 ± 5.36**	**4**	**0.06 ± 0.01**	**11.69 ± 7.41**
	FSCA-IS^c (q=1)	100.00%	6.99 ± 7.67	5	0.06 ± 0.01	11.25 ± 6.39
	FSCA-IS^c (q=2)	100.00%	5.83 ± 4.92	5	0.06 ± 0.01	11.69 ± 7.15
	FSCA-IS^c (q=3)	100.00%	5.25 ± 4.60	4	0.06 ± 0.01	11.69 ± 7.72
	SDS	**100.00%**	**4.61 ± 3.97**	**4**	**0.09 ± 0.01**	**14.11 ± 7.82**

conductor manufacturing," *IEEE Transactions on Automation Science and Engineering*, vol. 14, no. 2, pp. 894–904, 2017.

[14] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[15] C.-F. Chien and S.-C. Chuang, "A framework for root cause detection of sub-batch processing system for semiconductor manufacturing big data analytics," *IEEE Transactions on semiconductor Manufacturing*, vol. 27, no. 4, pp. 475–488, 2014.

[16] M. Maggipinto, C. Masiero, A. Beghi, and G. A. Susto, "A convolutional autoencoder approach for feature extraction in virtual metrology," *Procedia Manufacturing*, vol. 17, pp. 126–133, 2018.

[17] J. Wan, S. Pampuri, P. G. O'Hara, A. B. Johnston, and S. McLoone, "On regression methods for virtual metrology in semiconductor manufacturing," in *IEEE Conference on Automation Science and Engineering (CASE)*. IET, 2014.

[18] G. A. Susto and A. Beghi, "Least angle regression for semiconductor manufacturing modeling," in *Control Applications (CCA), 2012 IEEE International Conference on*. IEEE, 2012, pp. 658–663.

[19] A. T. McCray, J. McNames, and D. Abercrombie, "Locating disturbances in semiconductor manufacturing with stepwise regression," *IEEE Transactions on Semiconductor Manufacturing*, vol. 18, no. 3, pp. 458–468, 2005.

[20] P. Mitra, C. Murthy, and S. K. Pal, "Unsupervised feature selection using feature similarity," *IEEE transactions on pattern analysis and machine intelligence*, vol. 24, no. 3, pp. 301–312, 2002.

[21] Y. Cui and J. G. Dy, "Orthogonal principal feature selection," Department of Electrical and Computer Engineering, Northeastern University, Boston, Tech. Rep., 2008.

[22] J. Ranieri, A. Chebira, and M. Vetterli, "Near-optimal sensor placement for linear inverse problems," *IEEE Transactions on signal processing*, vol. 62, no. 5, pp. 1135–1146, 2014.

[23] F. Zocco and S. McLoone, "Mean squared error vs. frame potential for unsupervised variable selection," in *Intelligent Computing, Networked Control, and Their Engineering Applications*. Springer, 2017, pp. 353–362.

[24] P. Prakash, B. Honari, A. Johnston, and S. McLoone, "Optimal wafer site selection using forward selection component analysis," in *Advanced Semiconductor Manufacturing Conference (ASMC), 2012 23rd Annual SEMI*. IEEE, 2012, pp. 91–96.

[25] L. Puggini and S. Mcloone, "Forward selection component analysis: Algorithms and applications," *IEEE transactions on pattern analysis and machine intelligence*, 2017.

[26] A. Krause, A. Singh, and C. Guestrin, "Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies," *Journal of Machine Learning Research*, vol. 9, no. Feb, pp. 235–284, 2008.

[27] S. McLoone, A. Johnston, and G. Susto, "A methodology for efficient dynamic spatial sampling and reconstruction of wafer profiles," vol. 15, no. 4, pp. 1692 – 1703, 2018.

[28] J. Quiñonero-Candela and C. E. Rasmussen, "A unifying view of sparse approximate gaussian process regression," *Journal of Machine Learning Research*, vol. 6, no. Dec, pp. 1939–1959, 2005.

[29] S. McLoone, F. Zocco, M. Maggipinto, and G. A. Susto, "On optimising spatial sampling plans for wafer profile reconstruction," *IFAC-PapersOnLine*, vol. 51, no. 10, pp. 115 – 120, 2018, 3rd IFAC Conference on Embedded Systems, Computational Intelligence and Telematics in Control CESCIT 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2405896318305664

[30] B. M. Haddad, S. Yang, L. J. Karam, J. Ye, N. S. Patel, and M. W. Braun, "Multifeature, sparse-based approach for defects detection and classification in semiconductor units," *IEEE Transactions on Automation Science and Engineering*, vol. 15, no. 1, pp. 145–159, 2018.

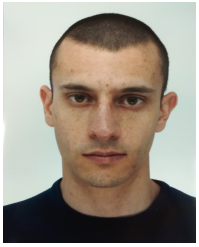
[31] A. Mandal, C. Jeff Wu, and K. Johnson, "Selc: Sequential elimination of level combinations by means of modified genetic algorithms," *Technometrics*, vol. 48, no. 2, pp. 273–283, 2006.

[32] T. J. Mitchell, "An algorithm for the construction of "d-optimal" experimental designs," *Technometrics*, vol. 16, no. 2, pp. 203–210, 1974.

[33] R. D. Cook and C. J. Nachtrheim, "A comparison of algorithms for constructing exact d-optimal designs," *Technometrics*, vol. 22, no. 3, pp. 315–324, 1980.
[34] Y. Huang, S. G. Gilmour, K. Mylona, and P. Goos, "Optimal design of experiments for non-linear response surface models," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 2018.
[35] D. R. Bingham and H. A. Chipman, "Incorporating prior information in optimal design for model selection," *Technometrics*, vol. 49, no. 2, pp. 155–163, 2007.

Gian Antonio Susto (S'11-M'16) received the M.S. degree (cum laude) in control systems engineering and the Ph.D. in Information Engineering from the University of Padova, Padova, Italy, in 2009 and 2013. He is currently Assistant Professor at University of Padova and founder at Statwolf Ltd. He has been a visiting student at the University of California, San Diego (2008-09), and at National University of Ireland Maynooth (NUIM) (2012), an Intern Researcher at Infineon Technologies Austria AG, Villach, Austria (2011) and post-doctoral associate at NUIM (2013). During his career he has been awarded the IEEE-CASE Best Student Conference Paper Award (2011), IEEE/SEMI-ASMC Best Student Paper Award (2012) and IEEE-MSC Best Student Paper Award (2012). He is Associate Editor for the IEEE Transactions on Semiconductor Manufacturing. He is the President of the Ethical Committee of the Human-Inspired Technology Center of the University of Padova. His research interests include manufacturing data analytics, machine learning, gesture recognition and partial differential equations control.

Marco Maggipinto received the M.S. degree (summa cum laude) in Telecommunications Engineering from the University of Padova, Padova, Italy in 2016. He is currently a Ph.D. Student at University of Padova. His research interests include deep learning, machine learning and computer vision. He has been an Erasmus student at the Universitat Politécnica de Catalunya (UPC)- Barcelona, Spain and a visiting researcher at the Queen's University of Belfast (QUB)- Belfast, UK and at Infineon Technology AG, Munich, Germany.



Federico Zocco received the B.S. degree in Mechanical Engineering and the M.S. degree in Robotics and Automation Engineering both from the University of Pisa, Italy, in 2013 and 2016, respectively. In 2013 he performed the B.S. thesis at the research center "E. Piaggio" of the University of Pisa. Currently he is a Ph.D. student at Queen's University Belfast, UK, and his main research interests are in theoretical and applied machine learning, in particular deep learning.

Seán McLoone (S'94-M'96-SM'02) received the M.Eng. degree (Hons.) in Electrical and Electronic Engineering and the Ph.D. degree in Control Engineering from Queen's University Belfast (QUB), Belfast, U.K., in 1992 and 1996, respectively. He was a Post-Doctoral Research Fellow, from 1996 to 1997, and a Lecturer, from 1998 to 2002, with QUB. He joined the Department of Electronic Engineering, NUI Maynooth, Maynooth, Ireland, in 2002, where he served as a Senior Lecturer, from 2005 to 2012, and as the Head of Department, from 2009 to 2012, before returning to QUB in 2013 to take up his current post as a Professor and the Director of the Energy Power and Intelligent Control Research Cluster in the School of Electronics, Electrical Engineering, and Computer Science. His research interests are in the general area of intelligent systems, with a particular focus on data based modeling and analysis of dynamical systems. This encompasses techniques ranging from classical system identification, fault diagnosis and statistical process control to modern computational intelligence based adaptive learning algorithms and optimization techniques. His research has a strong application focus, with many projects undertaken in collaboration with industry in areas such as process monitoring, control and optimization, time-series prediction, and inline sensor characterization. Prof. McLoone is a Chartered Engineer and a Fellow of the Institution of Engineering and Technology. He is a Past Chairman of the U.K. and Republic of Ireland Section of the IEEE.